

DOCUMENT RESUME

ED 410 283

TM 027 106

AUTHOR Reckase, Mark D.
TITLE Statistical Test Specifications for Performance Assessments:
Is This an Oxymoron?
PUB DATE Mar 97
NOTE 16p.; Paper presented at the Annual Meeting of the National
Council on Measurement in Education (Chicago, IL, March
25-27, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Assessment; *Generalizability Theory; Interrater
Reliability; *Performance Based Assessment; *Statistical
Analysis; Test Construction; Test Reliability; Test Use;
Test Validity
IDENTIFIERS High Stakes Tests; *Test Specifications

ABSTRACT

This paper argues that special procedures for constructing assessment tools containing performance assessment tasks are unnecessary and that current test methodology can easily be generalized to complex performance assessment tasks without destroying the desirable characteristics of those tasks. Reasonable statistical requirements for sound performance assessments can be described based on current experience in rater reliability, test reliability, generalizability, and validity. Content specifications are not the focus of this paper, but it is apparent that there is significant variation in the functioning of assessment tasks, and that content must be matched to objectives of the assessment. Considering performance assessment tasks as the target of instruction provides an appealing and straightforward model for assessment, but generalizing to other tasks is an issue that cannot be ignored. Two options are available to the test developer wishing to produce a performance assessment with generalizable results. The first is to select performance assessment tasks that are at least moderately intercorrelated, and the second is to increase the number of tasks administered until the desired level of generalizability is attained. Domain coverage and high stakes test use are other issues that must be explored. It is argued that statistical specifications such as inter-rater reliability, inter-task correlations, and generalizability coefficients are an important part of the design of performance assessments. (Contains 20 references.) (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Mark Reckase

**Statistical Test Specifications
for Performance Assessments: Is this an Oxymoron?¹**

Mark D. Reckase
ACT, Inc.

☒ This document has been reproduced as
received from the person or organization
originating it

☐ Minor changes have been made to
improve reproduction quality

☐ Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy

ED 410 283

In the period of time since the early development of educational achievement tests (e.g., Rice, 1902), great strides have been made in the process for developing tests. Procedures have been derived for developing tests to assess well defined constructs and to meet specific statistical requirements. Reckase (1996) provides an overview of the refinements in test development procedures since the 1930s and the current state of the art in standardized test development. Generally, test construction methods are now closer to a science than an art, with the desired statistical characteristics of tests well within the control of the test constructor.

Given the progress that has been made in standardized test development, it is somewhat surprising that some educators are suggesting that the psychometric underpinnings of the educational measurement process are no longer desirable, and that new methods for judging the quality of the currently popular performance assessment techniques are needed. For example, Gipps (1994) indicates "We do not ... see assessment as a scientific, objective, activity, this we now understand to be spurious" (p. 167). And, "Evaluation within the constructivist and naturalistic paradigms rejects the traditional criteria of reliability, validity

¹Paper presented at the annual meeting of the National Council on Measurement in Education and the American Educational Research Association, Chicago, March, 1997.

TM627106

and generalizability and looks instead for qualities such as trustworthiness and authenticity" (p. 168).

A somewhat less extreme position is presented by Moss (1992) who quotes a personnel communication from Allan Collins: "Collins notes that they [Frederiksen and Collins] have moved away from a sampling model of measurement to a performance model (similar to that used in the Olympic Games), where the quality of the performance and the fairness of the scoring are crucial but where replicability and generalizability of the performance are not" (p. 250). If such statements about the statistical indicators of test quality are taken at face value, then the construction of performance assessments should follow quite different rules of development than current standardized tests. Also, the results of the test development process should meet quite different standards of quality than other types of tests. Although it may be somewhat of an exaggeration of their positions, the above authors might consider the use of the terms "assessment" and "statistical specifications" in the same phrase to be an oxymoron. Of course, I disagree.

The purpose of this paper is to argue that "special" procedures for constructing assessment tools containing performance assessment tasks are unnecessary and that current test development methodology can easily be generalized to complex performance assessment tasks without destroying the desirable characteristics of those tasks. In particular, reasonable statistical requirements for sound performance assessments can be described based on current experience in the areas of (1) rater reliability, (2) test reliability, (3) generalizability, and (4)

validity. In addition, formal content specifications can be provided for performance assessments as well.

Specifications for Performance Assessments

Determining the skills and knowledge possessed by students using performance assessments represents both new and old methodology. Galton (1892) describes a performance assessment examination process for students in mathematics at Cambridge University at the turn of the century. "The examination lasts five and a half hours a day for eight days. All the answers are carefully marked by the examiners, who add up the marks at the end and range the candidates in strict order of merit" (p. 15). No doubt such procedures had been used at Cambridge University for some time.

Performance assessment procedures are considered innovative at this time because of the dominance of multiple-choice tests for large scale assessments since the 1930s. Certainly, performance assessments have been widely used in classrooms even when multiple-choice standardized tests were considered state of the art (I remember answering many extended response questions when I was in high school). The innovation is that performance assessment tasks are now appearing in large scale assessments (e.g., Breland, 1996; NAEP, 1994; Patience & Swartz, 1987), something that had not been the case since the initiation of wide spread standardized testing. Because of the large scale use of performance assessments

and the importance of the judgements made based on scores from such assessments, substantial amounts of research have been done on the qualities of performance assessments during the past few years.

Based on the research that is reported in the literature, and on personal experience at ACT, the following general recommendations are made for statistical test specifications for large scale performance assessments. These recommendations will be presented for typical uses for the assessments. The positions of Collins and Gipps will be considered in the discussion. My views and those of these authors are not necessarily in conflict. A common point of view may be derived through agreement on purposes.

Content Specifications

Although this paper is not about content specifications for performance assessments, content specifications will be discussed briefly to set the stage for the statistical specifications. There is an unjustified implication in some performance assessment literature that face validity is sufficient to show that complex performance assessment tasks are appropriate for use (e.g., Lesh & Lamon, 1992). Fortunately, research is beginning to appear that shows performance assessments are not automatically of high quality and that there is significant variation in the functioning of assessment tasks. For example, in a study of the use of performance assessment in science, Sugrue, Valdes, Schlackman, and Webb (1996) found substantial variation in the way that different types of items performed. They conclude that

"on the basis of the analysis of *p*-values and correlations presented here, one gets a sense of the sensitivity of performance to variation in format and context of items. Not only does performance vary depending on how we ask students to display their knowledge, but the type and extent of variation in performance is not consistent across types or levels of knowledge" (p. 15). Clearly, more work needs to be done to determine how to select tasks for performance assessment instruments.

Fortunately, sound guidelines for the construction of performance assessments are beginning to appear. Herman, Aschbacher, and Winters (1992) have produced a very usable guidebook for persons involved in the construction of performance assessments. They provide fairly clear guidelines for selecting assessment tasks that will match the objectives of the assessment. With more practical experience in the development of performance assessments, such guides will likely become more widely available.

Instructional Support

Some proponents of performance assessment consider facilitating learning as the most important function of educational assessment. Resnick and Resnick (1996) are clear supporters of this position.

"Learning improvement, we argue, is the first of the four reasons for which schools seek to measure the work of their student. Although the other reasons

include certification, accountability, and monitoring, for us and for increasing numbers of parents and lay people, learning improvement has the highest and most urgent priority" (p. 29).

If learning improvement is the use that is proposed for a performance assessment, what technical qualities should it have to support this goal?

This question can be answered in two different ways depending on the importance placed on using assessments to motivate students. Strictly from an instructional perspective, assessments used for instructional support should provide rich activities that match the goals of instruction and feedback to the student about the accomplishment of the goals. It is important that the student know that someone is paying attention to what they do. For these purposes, perhaps all that is needed is feedback from a credible source. That is, no technical requirements need be met since the assessment is part of the teacher/student interaction and individuals outside that interaction do not need to interpret the results.

However, suppose that the assessment must be high stakes to motivate the student to perform at their best. Resnick and Resnick (1966) indicate that "Without incentives for students to engage in the kind of challenging work that complex tasks represent at any grade level, it is unlikely that direct measures for assessment will fully produce the desired effect on learning" (p. 32). For these conditions to be met, the technical qualities of the assessment must also support high stakes use. The technical requirements for such use will be discussed

in a later section. However, they require full attention to the requirement in the *Standards for Educational and Psychological Testing* (1985).

The Target of Instruction

If the performance model of the Olympic Games presented above (Moss, 1922) is taken at face value, the goal of instruction becomes a high level of performance on the assessment task in the same sense that the goal for Olympic athletes is to win their event. All training is focused on improving the likelihood of achieving that goal. Under this model, the technical requirement of the assessment is reliable scoring. As with Olympic scoring for diving or ice skating, the requirements for the task are well known by all participants in advance, the judges are well trained on very specific rubrics, multiple judges are used, and the high and low judgements may be dropped to stabilize the averages of the judges ratings.

Generalizability to a larger universe of tasks or to other performances of the same task is not needed under this model of assessment because performance on the task is the goal. Nor is a high correlation with other measures. While, precision in scoring is clearly needed, whether spread in scores is required depends on whether a mastery model or an individual differences model is used for the assessment.

A mastery model implies that the percent of exact agreement is the statistic of choice for evaluating the quality of the assessment. Shifts in level of rating from one judge to the

next are unacceptable, even if the correlation between ratings is 1.0. However, if detecting differences in level of performance of students is critical, as it is in the Olympic Games analogy, then it is important that there be sufficient spread of scores to allow relatively fine distinctions in performance to be made. That is, performance assessment tasks and scoring rubrics should be designed so that score distributions on the tasks include all score categories.

The amount of spread of scores that is desirable is dependent on the inter-rater reliability that can be attained. Reckase (1996) showed that a very flat distribution of ratings is needed to support making fine distinctions in performance if the inter-rater reliability is fairly low (e.g., .5). As the inter-rater reliability increases, of course, the observed distribution is a closer match to the true score distribution. Inter-rater reliabilities of .7 or higher are now fairly common in the literature on performance assessment (see Brennan, 1996), so .7 would seem to be a reasonable target for correlations between pairs of ratings on a well constructed assessment. With a .7 inter-rater reliability, it is still important that substantial numbers of scores occur at the extremes of the score scales.

Generalizable Performance

While considering performance assessment tasks as the target of instruction provides an appealing and straightforward model for assessment, most psychometricians (e.g., Brennan, 1996) and educators (e.g., Herman, Aschbacher & Winters, 1992) would like to interpret the assessment task as representative of a more general set of tasks. That is, rather than focus on

the performance on the single assessment task, the goal is to infer that students could perform approximately equally well on other tasks of the same type.

Many researchers have shown that generalizing to other tasks from most performance assessments is questionable (Brennan, 1996; Dunbar, Koretz & Hoover, 1991; Shavelson, Baxter & Gao, 1993). Performance on one performance assessment task seems to be fairly idiosyncratic. Generalizability coefficients tend to be low, and trying to raise them by improving the rating process (by using more judges) seems to be relatively ineffective.

Two options are available to the test developer that desires to produce a performance assessment that yields generalizable results. The first is to select performance assessment tasks that are at least moderately intercorrelated. To achieve that goal, good inter-rater reliability will be required, and it may be necessary to pre-test a number of tasks so that those that intercorrelate can be selected for the assessment. This approach has the potential disadvantage that the task selection process might narrow the domain that is being assessed.

The second alternative is to increase the number of assessment task administered until the desired level of generalizability is attained. The number of tasks that will be required is dependent on the magnitude of the interitem correlations and the inter-rater reliability. If both of those statistical measures are low, many tasks will be needed to obtain a generalizable set.

Domain Coverage

A high generalizability coefficient means that students will likely be equally capable on other tasks of the same type. One can imagine a domain of tasks that are like the tasks that are already on the performance assessment, and student performance will likely generalize to all of those tasks. However, high generalizability for a particular performance assessment does not necessarily indicate that the level of performance will generalize to an entire domain. The tasks on the assessment must also span the domain. Messick (1996) has indicated that domain coverage is critical to the construction of performance assessments. "The concern that a performance assessment should provide representative coverage of the content and processes of the construct domain is meant to insure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly" (p. 10). For example, if the tasks on a performance assessment required students to create large volume containers from pieces of cardboard, the results would likely generalize to other problems of the same type, but not to tasks sampled from a broad domain of geometry tasks.

To provide generalizability to a full domain, the tasks must be shown to span the full domain and there must be high generalizability as well. If the domain is thought to be fairly unidimensional, defining a continuum of skills, then domain coverage can be demonstrated by showing that the assessment tasks provide information over the range of the continuum that is of interest. Tasks with an appropriate range of difficulty will be needed and they must correlate well with the composite score on the assessment instrument. If a complex

achievement domain is considered, than the tasks must be a representative sample from the domain and assessment should have a high generalizability coefficient. As the inferential leap needed to interpret the score increases, the psychometric support for that leap increases.

High Stakes with Multiple Uses

Many standardized tests have multiple uses, some of which may have important consequences for individuals. For example, the ACT Assessment is used for academic counseling, college planning, college placement, selection of students for admissions and scholarship, screening for athletic competition, etc. (ACT, 1996). To support these multiple uses, the test needs to have high reliability, high generalizability, documented construct validity, and evidence of predictive validity for specific uses. Performance assessments that have similar multiple uses and that provide evidence for important decisions that affect individuals need similar documentation to support the uses.

High stakes uses for performance assessment require all of the qualities of all of the previous uses. The scoring of the assessments must be reliable; the tests must be generalizable, either through high task correlations or sufficient number of tasks to adequately represent the domain; construct validity must be supported either through scale formation or evidence of domain coverage, and predictive validity needs to be demonstrated for some uses. The full psychometric requirements for high-stakes, multiple uses are detailed in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) so they

will not be repeated here. However, it is clear that statistical/psychometric indications of quality are a necessary component of performance assessments unless they are used solely as instructional tools. Thus, using statistical specifications and performance assessment together is not an oxymoron.

Conclusion

The goal of this paper has been to provide an argument that statistical specifications such as desired levels of inter-rater reliability, inter-task correlations, and generalizability coefficients are an important part of the design of performance assessments. Yet, few texts on performance assessment even mention these terms. For example, *Testing for Learning* (Mitchell, 1992) does not even include the term "reliability" in the index.

The lack of discussion of statistical specifications for performance assessments is not a result of lack of research in the area. Numerous studies and articles have been published on the topic, some of which have been referenced in this paper. Rather, proponents of performance assessment suggest that a different theoretical framework is needed for evaluating performance assessments. The position taken here is that there may be some merit for the position if the performance assessments are used solely as instructional tasks, but even that position is questionable if high stakes uses for the assessment results are needed to motivate the students. For all other uses, some statistical requirements are needed to support

the desired use. To the extent possible in brief this paper, the types of statistical requirements that are important have been listed. A more thorough presentation on statistical test specifications for performance assessments that summarizes the collective experience of practitioners in the field is clearly needed. Perhaps next year

References

- ACT (1996). *ACT assessment user handbook*. Iowa City, IA: ACT, Inc.
- AERA, APA, & NCME (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Breland, H. M. (1996, April). *Writing skill assessment: Problems and prospects*. Princeton, NJ: Educational Testing Service.
- Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: U.S. Department of Education.
- Dunbar, S. B., Koretz, D. M. & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4, 289-303.
- Galton, F. (1892). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan.
- Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: The Falmer Press.
- Herman, J. L., Aschbacher, P. R. & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Lesh, R. & Lamon, S. J. (1992). *Assessment of authentic performance in school mathematics*. Washington, DC: AAAS Press.
- Messick, S. (1996). Validity of performance assessment. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: U.S. Department of Education.
- Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools*. New York: The Free Press.
- Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research*, 62(3), 229-258.
- NAEP (1994). *NAEP 1992 writing report card*. Princeton, NJ: Educational Testing Service.

- Patience, W. & Swartz, R. (1987, April). *Essay score reliability: Issues in and methods of reporting the GED Writing Skills Test scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.
- Reckase, M. D. (1996). Test construction in the 1990s: Recent approaches every psychologist should know. *Psychological Assessment*, 8(4), 354-359.
- Reckase, M. D. (1996, June). *Desirable characteristics of score distributions on performance assessment tasks*. Paper presented at the annual meeting of the Psychometric Society, Banff, Canada.
- Resnick, D. P. & Resnick, L. B. (1996). Performance assessment and the multiple functions of educational measurement. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems, and challenges*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Rice, J. M. (1902). Educational research: A test in arithmetic. *The Forum*, 34, 281-297.
- Shavelson, R. J., Baxter, G. P. & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215-232.
- Sugrue, B., Valdes, R., Schlackman, J. & Webb, N. (1996, March). *Patterns of performance across different types of items measuring knowledge of Ohm's Law* (Technical Report 405). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.